AUTOMATIC LABELLING OF GERMAN PROSODY

Stefan Rapp[*]

Institut für Maschinelle Sprachverarbeitung (IMS),
Universität Stuttgart, Azenbergstr. 12, 70174 Stuttgart, Germany
e–mail: rapp@sony.de

ABSTRACT

One limitation in prosody research is the lack of sufficient prosodically labelled speech data. In this paper, we present research on an automatic labelling system that is able to produce a phonological tonal labelling according to the ToBI like intonation model for German developed by Féry. The system is not totally dependent on the specific language and/or labelling system, as it uses corpus based techniques such as an HMM based word, syllable and phoneme segmentation and a decision tree learning algorithm (C4.5) for the phonetic-phonology mapping. The current system was trained on about 1 hour of expert prosodically labelled speech from a single male radio news announcer. We present experiments for finding a suitable feature set drawn from features that describe the prosodic correlates fundamental frequency, duration and intensity as well as some lexical and syntactic features. With the best feature set, we achieve a recognition rate of 78.7% for speaker dependent recognition of ToBI labels (simultaneously predicting prominence and phrasing) and 86.9% for the simpler accented/not accented decision. Although the system's accuracy is well below that of human transcribers, it is a useful tool actively used in our laboratory due to it's ability to process large amounts of speech data at low costs.

1. INTRODUCTION

At present, we do not fully understand the meaning speakers encode with prosody. In order to better understand the role pitch accents and phrase boundaries play in communicating meaning from a speaker to a hearer, we need prosodically annotated speech corpora at an adequate level of abstraction so that we can check and find hypotheses about how prosody relates to syntactic structure, semantics and pragmatics. In our opinion, the tone sequence model of intonation [2] respectively the German counterpart we use in our research [4, 6] is sufficiently abstract. However, building up a database with a tonal labelling is very much work, and, as it requires phonetic expertise, rather expensive. Opposed to segmental labelling, where by now automatic labelling procedures building on speech technology are approaching human labelling accuracy and are able to produce huge amounts of segmentally labelled speech data, fewer research was done in automatic prosody labelling [13, 11, 3]. In this paper, we present an automatic procedure that produces a phonological tonal labelling according to the ToBI like intonation model for German developed by Féry [4, 6]. Disregarding downstep, the tonal tier of this labelling system consists of five underlying pitch accents, L*H, H*L, L*HL, HH*L and H*M as well as four further labels, L*, H*, ..H and ..L that are available in the surface structure that the prosodic labelling describes. For marking prosodic phrase structure, the system has four boundary tones, %, H%, L% and -.[1] As both a pitch accent label and a phrase boundary label can be associated with the same syllable, the task is to predict pairs $p \in ((\textit{pitch accent labels} \cup \{\{\}\}) \times (\textit{boundary tone labels} \cup \{\{\}\}))$ for every syllable.

2. ARCHITECTURE

Conceptually, the system consists of three layers, acoustic, phonetic, and phonological. The acoustic layer deals with continuous data, while for both the phonetic and phonological layers, the syllable is the basic unit of measurement. Figure 1 gives an overview about the system's components.

The elements of the phonetic layer, right in the middle of fig. 1, are derived from the acoustic layer by several system components. First, the automatic segmentation system Alphons, which is based on HTK and the German part of the CELEX lexicon [1], segments the speech data into words, syllables and phonemes [8, 10]. From these segmentations, duration parameters are extracted as the first group of features in the phonetic layer. **Dur-DistToNextP** gives the distance (in seconds) to the next detected speech pause, the length of it is given in **Dur-LenOfNextP**. From a statistic of phoneme lengths **Dur-SylLenExpected** is calculated by summing the mean durations over all the phonemes that are contained in a syllable. Further to that the actual syllable duration from the syllable

[*] The author is now with Sony International (Europe) GmbH, European R&D, Adv. Developments, Stuttgarter Str. 106, 70736 Fellbach, Germany.

[1] L% and - are not included in Féry's analysis and were introduced into the system by Mayer [6].

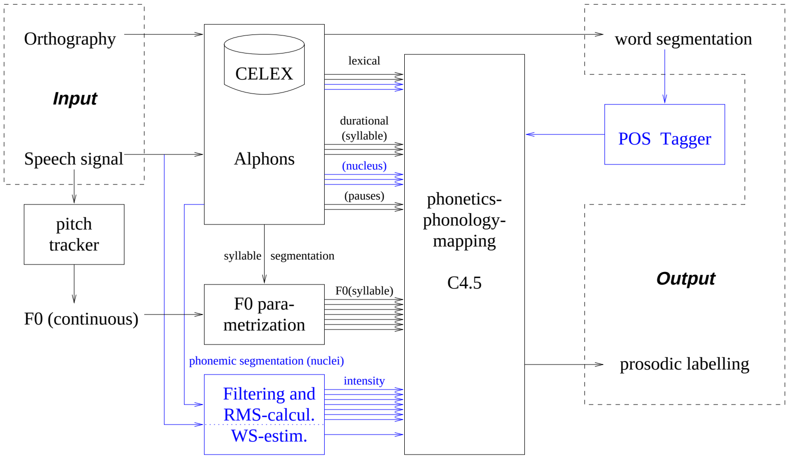

Figure 1: An overview of the system.

segmentation is contained in **Dur-SylLenMeasure**. **Dur-SylLenRelative** is the ratio of the previous two attributes and indicates the degree to which a syllable is longer or shorter as expected. Similar features were also calculated for the syllable nucleus instead of the syllable. **Dur-NucLenExpected** gives the mean duration of the syllable's nucleus, **Dur-NucLenMeasure** the actually occurring length and **Dur-NucLenRelative**, again, their ratio.

Second, the fundamental frequency estimated by the ESPS pitch tracker is parameterized into 7 phonetically interpretable parameters per syllable [9], utilizing the syllable segmentation of Alphons. They describe the pitch movement in a two syllable window and add to the durational phonetic features. The F_0 of a two syllable window is approximated by a superposition of three functions, a $\tanh(x)$, an e^{-x^2} and a constant function. The seven parameters that describe the approximating superposition of functions can be algorithmically extracted from the F_0 estimated by the ESPS pitch tracker. **F_0-TonalDiff** correlates with the height of a rise or fall, **F_0-TonalSteep** with the steepness of a rise or fall, **F_0-TonalAlign** with the position at which a rise or fall occurs, **F_0-PeakHeight** with the height of a peak (or the deepness of a valley), **F_0-PeakSteep** with its steepness, **F_0-PeakAlign** with its temporal alignment, and **F_0-Level** with the overall F_0 level.

Eight features describing the intensity are estimated by also referring to a segmentation of Alphons. The phonemic segmentation is used to find the syllable nuclei portions of the signal, on which all intensity estimations are based. **Int-RMS0-8k** gives the mean energy over the syllable's nucleus. Overall energy was not adjusted from news story to news story. **Int-RMS2-8k** and **Int-RMS4-8k** were used for a description of spectral tilt, they were normalized wrt. overall energy to give the amount of energy found in the band above 2kHz or 4kHz respectively. The remaining **Int-RMS0-500**, **Int-RMS500-1k**, **Int-RMS1-2k** and **Int-RMS2-4k** complement the energy contained in individual sub-bands of the spectrum and were also normalized to the overall energy. As a last feature to describe intensity, **Int-WordStressGuess** was used. It is the binary output of a classifier trained with C4.5 on the task of predicting lexical word stress from 12 Mel frequency cepstral coefficients taken from the center of the nucleus vowel[2] [10]. On a different corpus, this classifier approaches performance of another classifier predicting word stress from the two most important correlates of spectral tilt, 'skewness' and 'rate of closure' for which the data was extracted from speech signals by a phonetic expert [10].

From the features of the phonetic layer, a phonological description of intonation, i.e. ToBI-Labels, is predicted by a decision tree. In addition to the phonetic features, the classifier is given access to four lexical features also produced by Alphons and one syntactic feature. **Lex-NucType** gives a classification of the nucleus type into short, long, diphthong or schwa, **Lex-NucVowel** gives the phoneme of the nucleus. **Lex-WordStress** is a binary feature that is true for syllables with associated word stress, for words not contained in the lexicon (Alphons generates pronunciations for these with grapheme to phoneme conversion rules), no syllable has assigned word stress. The position of the syllable in the word is determined by **Lex-Syls2WordEnd** as the number of syllables that follow up to the next word boundary. As the only syntactical information, part of speech tags are produced by Schmid's part of speech tagger [12] and are included in the **Syn-PartOfSpeech** attribute.

The decision tree for the phonetic-phonology-mapping is generated automatically from a database of expert prosodically labelled speech by C4.5 [7]. The use of corpus based

[2] the point that is equidistant to both nucleus start and end, not the point where maximum intensity is reached

techniques such as the HMM based word, syllable and phoneme segmentation and the decision tree learning algorithm makes the presented system not totally dependent on the specific language and/or labelling system used. However, the approach requires the existence of training data. For our experiments, we use more than one hour of professionally read, real life German radio news stories by a single male announcer. It was gathered from Deutschlandfunk via Digital Satellite Radio in very good quality. The speech material was transliterated, automatically word- and syllable-aligned and a full prosodic labelling according to [5] was performed. The available speech material was divided into a training and a test set consisting of 10445 resp. 2436 syllables and was kept constant for all runs.

3. EXPERIMENTS

We carried out a number of experiments to optimize recognition performance by adding or removing acoustic, syntactic or lexical information. The investigated sets of features and the recognition rates are reported in table 1.

The results are interesting from a phonetic point of view. The major findings are that, at least for the given data, none of several intensity measures contributed to recognition accuracy, not even the features that try to represent spectral tilt (first group in table 1). As can be seen from the second group of experiments, this also holds when the durational features are missing. Also, including part-of-speech-tags as features did not improve recognition accuracy as the third group of table 1 shows. The next set of experiments show, that part-of-speech information certainly contributes, when we do not have access to acoustic measurements as it would be the case for prosody label generation for text-to-speech. Possibly as a consequence of the rather small amount of training data, adding features of surrounding syllables did not improve accuracy either which can be seen from the results of the fifth group. Finally, and in disagreement with previous findings, syllable duration measurements outperformed nucleus duration measurments. The results must be interpreted carefully because they are not based on a large population of speakers, and might depend on the way this specific database was collected.

4. CONCLUSIONS

The system achieves accuracy of 78.7% for speaker dependent recognition of ToBI labels (predicting accents and phrase boundaries simultaneously), and 86.9% for the simpler accented/not accented decision. The system is in active use in laboratory research due to it's ability to process large amounts of speech data at low costs, although the system's accuracy is below that of human labellers.

References

[1] R. H. Baayen, R. Piepenbrock, and H. van Rijn. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1993.

[2] M. E. Beckman and J. B. Pierrehumbert. Intonational structure in Japanese and English. *Phonology Yearbook*, 3:255–309, 1986.

[3] N. Campbell. Autolabelling Japanese ToBI. In *ICSLP96 Proceedings Fourth International Conference on Spoken Language Processing*, pages 2399–2402, October 1996.

[4] C. Féry. *German Intonational Patterns*. Niemeyer, Tübingen, 1993.

[5] J. Mayer. Transcribing German intonation — the Stuttgart system. Manuscript, Univ. Stuttgart, 1995.

[6] J. Mayer. *Intonation und Bedeutung. Aspekte der Prosodie-Semantik-Schittstelle im Deutschen*. Dissertation, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1997.

[7] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kauffmann, San Mateo, CA, 1992.

[8] S. Rapp. Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An aligner for German. In *Workshop "Integration of Language and Speech in Academia and Industry"*, Moscow, November 1995. ELSNET goes east and IMACS. http://www.ims.uni-stuttgart.de/˜rapp/aligner.ps.gz.

[9] S. Rapp. Goethe for prosody. In *ICSLP96 Proceedings Fourth International Conference on Spoken Language Processing*, pages 1636–1639, October 1996.

[10] S. Rapp. *Automatisierte Erstellung von Korpora für die Prosodieforschung*. PhD thesis, University of Stuttgart, Institut für maschinelle Sprachverarbeitung, 1998.

[11] K. N. Ross. *Modeling of Intonation for Speech Synthesis*. PhD thesis, Boston University College of Engineering, 1995.

[12] H. Schmid. Improvements in part-of-speech tagging with an application to German. In *Proceedings of EACL SIGDAT-Workshop*, Dublin, Ireland, 1995. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.ps.gz.

[13] C. W. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481, October 1994.

F_0-TonalDiff	F_0-TonalSteep	F_0-TonalAlign	F_0-PeakHeight	F_0-PeakSteep	F_0-PeakAlign	F_0-Level	Dur-DistToNextP	Dur-LenOfNextP	Dur-SylLenExpected	Dur-SylLenMeasure	Dur-SylLenRelative	Dur-NucLenExpected	Dur-NucLenMeasure	Dur-NucLenRelative	Int-RMS0-8k	Int-RMS2-8k	Int-RMS4-8k	Int-RMS0-500	Int-RMS500-1k	Int-RMS1-2k	Int-RMS2-4k	Int-WordStressGuess	Lex-NucType	Lex-NucVowel	Lex-WordStress	Lex-Syls2WordEnd	Syn-PartOfSpeech	*prepruning* = 2	*prepruning* = 5
○	○	○	○	○	○	○	○	○	○	○	○														○	○		76.85	78.65
○	○	○	○	○	○	○	○	○	○	○	○				○										○	○		76.89	78.28
○	○	○	○	○	○	○	○	○	○	○	○					○									○	○		76.68	78.04
○	○	○	○	○	○	○	○	○	○	○	○						○								○	○		76.23	77.67
○	○	○	○	○	○	○	○	○	○	○	○				○	○	○	○	○	○	○				○	○		74.75	76.97
○	○	○	○	○	○	○	○	○	○	○	○											○			○	○		76.52	78.12
○	○	○	○	○	○	○	○	○																	○	○		75.29	77.30
○	○	○	○	○	○	○	○	○							○										○	○		74.18	76.77
○	○	○	○	○	○	○	○	○								○									○	○		75.00	76.97
○	○	○	○	○	○	○	○	○									○								○	○		74.10	76.35
○	○	○	○	○	○	○	○	○							○	○	○	○	○	○	○				○	○		73.69	75.25
○	○	○	○	○	○	○	○	○														○			○	○		75.41	76.89
○	○	○	○	○	○	○	○	○	○	○	○														○	○		76.85	78.65
○	○	○	○	○	○	○	○	○	○	○	○														○	○	○	77.26	78.49
○	○	○	○	○	○	○	○	○	○	○	○														○	○	$\underset{\leftarrow}{\circ}$	75.90	78.00
○	○	○	○	○	○	○	○	○	○	○	○														○	○	$\vec{\circ}$	76.77	76.77
○	○	○	○	○	○	○	○	○	○	○	○														○	○	$\underset{\leftarrow}{\vec{\circ}}$	75.86	77.09
									$\underset{\leftarrow}{\vec{\circ}}$																$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	73.77	74.96
									○																○	○	$\underset{\leftarrow}{\vec{\circ}}$	75.12	73.77
									○																○	○	○	73.89	73.60
									○																○	○		72.66	73.32
																									○		$\underset{\leftarrow}{\vec{\circ}}$	73.15	73.73
																										○	$\underset{\leftarrow}{\vec{\circ}}$	73.36	73.40
○	○	○	○	○	○	○	○	○	○	○	○														○	○		76.85	78.65
$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$														$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$		75.33	76.89
$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$														$\vec{\circ}$	$\vec{\circ}$		75.45	76.60
$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$														$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$		75.62	76.68
$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	○	○	○	○	○														○	○		75.53	77.13
$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$	○	○	○	○	○														○	○		73.97	78.04
$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	○	○	○	○	○														○	○		74.22	76.48
○	○	○	○	○	○	○	○	○	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$														○	○		75.99	77.30
○	○	○	○	○	○	○	○	○	$\vec{\circ}$	$\vec{\circ}$	$\vec{\circ}$														○	○		75.21	77.05
○	○	○	○	○	○	○	○	○	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$														○	○		74.59	75.74
○	○	○	○	○	○	○	$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$	○	○	○														○	○		76.23	77.71
○	○	○	○	○	○	○	$\vec{\circ}$	$\vec{\circ}$	○	○	○														○	○		76.15	77.91
○	○	○	○	○	○	○	$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$	○	○	○														○	○		75.90	77.67
○	○	○	○	○	○	○	○	○	○	○	○														$\underset{\leftarrow}{\circ}$	$\underset{\leftarrow}{\circ}$		75.99	78.37
○	○	○	○	○	○	○	○	○	○	○	○														$\vec{\circ}$	$\vec{\circ}$		76.31	77.26
○	○	○	○	○	○	○	○	○	○	○	○														$\underset{\leftarrow}{\vec{\circ}}$	$\underset{\leftarrow}{\vec{\circ}}$		76.19	78.12
○	○	○	○	○	○	○	○	○	○	○	○														○	○		76.85	78.61
○	○	○	○	○	○	○	○	○				○	○	○											○	○		75.78	77.55
○	○	○	○	○	○	○	○	○	○	○	○												○		○	○		76.23	78.41
○	○	○	○	○	○	○	○	○				○	○	○									○		○	○		76.23	76.68
○	○	○	○	○	○	○	○	○	○	○	○													○	○	○		77.46	78.37
○	○	○	○	○	○	○	○	○				○	○	○										○	○	○		77.01	76.89

Table 1: Experimental results for various feature sets. Results are reported for two settings of the decision tree induction program. The first, which has prepruning set to 2 is the default setting of C4.5. With the second, by setting prepruning to 5, the algorithm is forced to generalize moderately stronger.

In the table, $\circ$ means that the feature of the syllable is available to the decision tree learning algorithm, $\underset{\leftarrow}{\circ}$ denotes that the same feature of the previous syllable is added to the set of available features as well. $\vec{\circ}$ represents cases in which the same feature of the next syllable is added, $\underset{\leftarrow}{\vec{\circ}}$ denotes that all three features, that of the current, previous and next syllable are available. In the case of the Syn-PartOfSpeech feature, the little arrows refer to the next and previous word instead of syllable.